Salience and pointing in multimodal reference

Conference or Workshop Item

In: Proceedings of PRE-CogSci 2009, 29 July 2009, Amsterdam, The Netherlands

Salience and Pointing in Multimodal Reference

Paul Piwek (P.Piwek@open.ac.uk)
Centre for Research in Computing, The Open University, Walton Hall
Milton Keynes, MK7 6AA United Kingdom

Abstract

Pointing combined with verbal referring is one of the most paradigmatic human multimodal behaviours. The aim of this paper is foundational: to uncover the central notions that are required for a computational model of human-generated multimodal referring acts. The paper draws on existing work on the generation of referring expressions and shows that in order to extend that work with pointing, the notion of salience needs to play a pivotal role. The paper investigates the role of salience in the generation of referring expressions and introduces a distinction between two opposing approaches: salience-first and salience-last accounts. The paper then argues that these differ not only in computational efficiency, as has been pointed out previously, but also lead to incompatible empirical predictions. The second half of the paper shows how a salience-first account nicely meshes with a range of existing empirical findings on multimodal reference. A novel account of the circumstances under which speakers choose to point is proposed that directly links salience with pointing. Finally, a multi-dimensional model of salience is proposed to flesh this model out.

Keywords: Generation of Referring Expressions; Multimodal Reference; Salience; Pointing Gestures; Deixis.

Introduction

Researchers on human pointing gestures have observed that pointing is essentially a means to "reorient the attention of another person so that an object becomes the shared focus of attention" (G. Butterworth, 2003). Somewhat surprisingly, this insight does not seem to have a counterpart in computational models of multimodal referring expression generation. In these accounts, *focus of attention*, *accessibility* and *salience*, three notions whose interrelationships we examine in more detail in the next section, are absent. Pointing is treated as either a fallback strategy for when verbal means fall short, or as expressing a property (i.e., as denoting a set of objects) in the same way that words, such as 'red' or 'bird', express properties.

For example, Lester, Voerman, Towns, and Callaway (1999) describe a system that only produces a pointing act, when a pronoun does not suffice to identify the target. Similarly, Claassen (1992) introduces an algorithm which only uses pointing if no purely verbal means of identification is possible, and Sluis and Krahmer (2001) describe an algorithm that only generates a pointing act if a purely verbal referring act becomes too complex. More recently, Krahmer and Sluis (2003) treat pointing acts not very different from words: as expressing a property. A pointing act identifies a subset of objects in the domain. Their algorithm assigns costs to the properties that are included in a referring expression. A graph-based algorithm is employed to find the cheapest combination of properties for referring to an object.

This is not to say that none of the models of referring expression generation and interpretation use notions such as attention, accessibility, or salience – a notion that will occupy a central place in the model that is offered in this paper. For example, visual salience plays a pivotal role in the interpretation and generation algorithms of Kelleher, Costello, and Genabith (2005). Similarly, Choumane and Siroux (2008) model visual salience for interpretation. Neither of these accounts do, however, directly relate salience to pointing gestures: Kelleher et al. (2005) only deals with verbal referring acts, whereas Choumane and Siroux (2008) view pointing acts rather narrowly as designating an object, rather than playing the dynamic role of changing the focus of attention.

The aim of this paper is to unpick the relation between salience and pointing and lay the foundations for a computational account based on this relation. The next section makes the assumptions behind the current approach explicit, and spells out the relation between the notions of salience, accessibility and focus of attention. Next, the role of salience in the generation of referring expressions is examined. We distinguish between two opposing approaches for dealing with salience: salience-first and salience-last accounts, and argue that these differ not only in computational efficiency, as has been pointed out previously, but also lead to diverging empirical predictions. The second half of the paper shows how a salience-first account nicely meshes with a range of existing empirical findings on multimodal reference. A novel account is put forward of the circumstances under which speakers choose to point. This account directly links salience with pointing. Finally, it is fleshed out by introducing a multi-dimensional model of salience for multimodal reference.

Assumptions and Terminology

The situations that we aim to model have three main ingredients: a speaker, an addressee and a visually shared domain of discourse. The speaker's goal (or intention) is to identify an object, the target, for the addressee in the domain of discourse. To achieve this goal, the speaker can use both language and pointing gestures. The scope of the model is restricted to cases in which the speaker is referring to objects in the visually shared domain and, if the speaker points, the target is among the objects that the speaker points at. This excludes cases such as those discussed by Clark, Schreuder, and Buttrick (1983) and Goodwin (2003). For example, Clark et al. (1983) discuss a speaker who says 'I worked for those people' whilst pointing at a newspaper. In this instance, the speaker referred to the publishers of the newspaper. Cases like this one, where the speaker refers to an object that is not in the visually shared domain and points at an object which is different from the target, are beyond the scope of the current study.

The aim of the model is two-fold: A) to produce expressions that are identical to those that humans produce in similar situations and B) to be a model that generates referring expressions using similar mechanisms as humans do. The emphasis is, however, on A: the model has been constructed using a range of findings on the expressions humans produce under various conditions. B is only addressed to the extent that we borrow notions from cognitive psychology, such as salience, to frame the model and make sure that the model is consistent with experimental results regarding the timing of speech and gestures. The model is *not* intended as an engineering solution to the generation of referring expressions. For example, the following are not aims in themselves: to generate the shortest expression that uniquely identifies the referent, generate an expression that uniquely identifies a referent in the computationally least costly way, or to produce expressions that are easiest for humans to comprehend.

The model is put forward as an information-processing model; it rests on the assumption that we can describe a cognitive activity in terms of the representations and processes, the computations, that are involved in that activity. We assume that, even though the human brain implements these computations, the nature of the activity can be characterized in terms of the computations only. See Ruiter (2000) for an excellent description of the information-processing approach to cognitive modelling, specifically, for the study of multimodel behaviours. Here, we would like to note that an information-processing approach does introduce considerations of computational cost: if we, as humans, can perform a particular task within certain temporal constraints, this does put constraints on the efficiency of the computational mechanisms that the model invokes.

The model that we take as a point of departure, the Incremental Algorithm (IA), was devised by Dale and Reiter (1995) to address some of the shortcomings of previous computational models for referring expression generation. Dale and Reiter (1995) argue that 'the simplest [model] may be the best, because it seems to be closest to what human speakers do.' In other words, their critique of prior models focuses specifically on their cognitive plausibility. They identified two specific weaknesses of these models: they generated expressions that would never be generated by human speakers and put unrealistic computational demands on the generator. The IA is compatible with one of the leading cognitive models of speech production, Levelt's blueprint for a speaker (Levelt, 1989). The IA has in common with the blueprint the assumption that generation starts from an intention. IA divides the generation task into the problems of *what to say* and *how to say it*, a division mirroring the distinction made in the blueprint between the conceptualizer and the formulator. The model that is proposed here is concerned primarily with the problem of what to say. In terms of the blueprint model, it focuses on the conceptualizer, the module which takes an intention and generates a preverbal message using various resources, such as a discourse model and situational knowledge.

Ruiter (2000) has proposed an extension to the blueprint for multimodal production. He suggests that the conceptualizer produces both a preverbal message (a specification of the information that has to be expressed by means of language) and, what he calls, a sketch (a specification of the information that has to be expressed by means of a gesture). The preverbal message and sketch are planned together in the conceptualizer. Ruiter (2000) also argues that subsequent processing stages operate mainly independently and in parallel: the preverbal message is send to a formulator and the sketch to a gesture planner. Synchronization is explained by assuming that the formulator is only activated once the gesture planner has constructed a motor plan for execution. Thus, the formulator produces a phonetic plan for execution only after the motor plan for the gesture is ready. This assumption accounts for the empirical finding that the onset of gestures precedes that of the accompanying speech (Levelt, Richardson, & Heij, 1995; Ruiter, 1998; Feyereisen, 2007).

The assumption that gesture and speech derive from a single starting point – the intention, goal or, in McNeill's terminology, growth point (McNeill, 2005) – is common to most psycholinguistic theories of language and gesture. They assume some sort of process which plans an initial specification of the multimodal act. Divergences relate to the degree of interaction between the language and gesture planning at later stages, with at least three distinct hypotheses: A) the Free Imagery hypothesis according to which gestures are constructed mainly independently of language (Krauss, Chen, & Chawla, 1996; Ruiter, 2000), B) the Lexical Semantics Hypothesis which says that gestures, specifically iconic ones, are generated from the semantics of lexical items (B. Butterworth & Hadar, 1989) and C) the Interface Hypothesis (Kita & Özyürek, 2003) according to which there exists a representation which mediates between both spatio-motoric and linguistic information. The current model does not take a side in this dispute. The model focuses on the initial production stages which, in de Ruiter's terms, are completed once a preverbal message and a sketch have been produced. Though the standard formulation of the IA does not take linguistic information into account, it is possible to integrate syntactic constraints, as demonstrated by Krahmer and Theune (2002). Our focus will be on the microstructure of the conceptualization processes. We aim to go beyond the level of detail common in information-processing theories, which are usually formulated at the level of box and arrow drawings. The formalization is meant to generate specific predictions that will hopefully give rise to new empirical studies.

As we already pointed out, our model applies to settings that include two participants (a speaker and an addressee) and a visually shared situation inhabited by discrete objects. Now suppose that we give our speaker and addressee the following task: each is to independently select an object *and* try to select the same object as the other participant. This is an instance of a Schelling task (Schelling, 1960). Remarkably,

even though our participants are not allowed to communicate, they are reasonably likely to succeed in selecting the same object. This is because, even though the participants are not allowed to communicate, they are bound to have some common ground as a result of various factors. Clark et al. (1983) mention shared experiences (e.g., the fact that they are looking at the same scene), previous communication (e.g., one of them might have referred to some object in the past), and shared community membership (e.g., they may both be Dutch nationals). We would like to add to this inventory similar perceptual and cognitive capabilities (e.g., perceiving some objects as more prominent, because of their size or colour). In short, relative to the common ground,[1] some objects will be more prominent/salient than others to both of them. In this paper, a notion of salience along these lines, best referred to as *joint* salience, plays a central role. We will formalize this notion of salience by associating numerical salience values with objects in the shared situation. The values represent the salience of the objects relative to the interlocutors' common ground. We also provide equations that describe how the salience values change as a result of verbal and non-verbal actions, following the notation of Theune (2000) and Krahmer and Theune (2002).

We have introduced salience in terms of the Schelling task and emphasized its dependence on the common ground. The notion is closely related to both accessibility and focus of attention. Accessibility is defined by Kahneman (2003) as: "[...] the ease (or effort) with which particular mental contents come to mind. The accessibility of a thought is determined jointly by the characteristics of the cognitive mechanisms that produce it and by the characteristics of the stimuli and events that evoke it. [...] the determinants of accessibility subsume the notions of stimulus salience, selective attention, specific training, associative activation, and priming."[2] The notion of a focus of attention[3] can be related to accessibility by characterizing the focus of attention at some point in time t as the set of most accessible objects at time t. Accessibility, focus of attention and salience are closely related, though our interpretation of salience has a common/shared dimension which is absent in the straightforward interpretations of accessibility and focus of attention. The latter two, as opposed to (joint) salience, are defined purely from the individual's point of view.

[1] For a detailed discussion of the notion of common ground, see Clark (1996) which dispells some of the misconceptions that have arisen about this notion. The notion of common ground is often associated with one specific psychologically implausible version, common ground iterated, which requires an infinitely large mental capacity. Other versions, such as common ground shared basis, do, however, not have this limitation and provide a sound logical basis for mental representations, as worked out in detail by Barwise (1989).

[2] A similar cognitive notion of accessibility, grounded in neural activation, has been advocated by Mira Ariel as way to model differences between various types of referring expressions, including pronouns, demonstratives and definite descriptions (Ariel, 1990).

[3] A notion which was poineered in Computational Linguistics by Grosz and Sidner (1986).

Salience: first or last?

Throughout this paper, the Incremental Algorithm (IA) as first proposed in Dale and Reiter (1995) is used as a starting point. The IA works on the assumption that there is a universe or domain of objects $\mathcal{U}$ which includes a target r, the object the speaker intends to refer to. In order to refer to r, the speaker constructs a preverbal description D consisting of a set of properties $P_1, \ldots, P_n$ such that the intersection of these properties equals $\{r\}$. In other words, the description is such that it uniquely identifies r. Note that D is preverbal; the IA does *not* decide how the preverbal description is expressed in language,[4] it only chooses the properties that need to be expressed. Each property is treated extensionally[5] as a subset of $\mathcal{U}$ and properties are organized as belonging to attributes (e.g., the properties *red*, *green*, ... are associated with the attribute *colour*). Attributes are ordered, where the ordering indicates which attributes are preferred for constructing a description.

The algorithm works as follows: it starts with the empty description $D = \emptyset$ and a context set C which is initialized with the domain: $C = \mathcal{U}$, and iterates through the ordered list of attributes. The algorithm fails if the end of the list is reached. On each iteration, the following steps are taken:

1. The best property P belonging to the current attribute is selected, i.e., the property P which has the smallest non-empty intersection with C and includes r.

2. **If** $C - P \neq \emptyset$ (where $C - P$ stands for the set of objects in C that are ruled out by P), **then**:
   $C = C \cap P$ and $D = D \cup \{P\}$

3. **If** $C = \{r\}$ **then**:
   return D, unless D includes no property from the top-ranked attribute, in which case add an appropriate property from this attribute to D and return the result.[6]

There are two principal ways to add salience to this account. They can be compared most easily by assuming that salience S_r is a property, i.e., a subset of $\mathcal{U}$ that can be computed if we know the salience value of each of the objects in $\mathcal{U}$ and the identity of the target r:

> S_r, the *salience property for r*, is the set of objects whose salience value is above some threshold value which is defined as the salience value of r minus a confidence interval (see Figure 1).

[4] That is, it does not decide whether a property is realized as a noun, adjective or adverb and also does not govern the choice of determiner. Choice of determiner involves deciding between, for example, 'the', 'this' and 'that'. See Piwek, Beun, and Cremers (2008) for an empirical study into this issue.

[5] In order to avoid notational clutter, we use P to refer both to the name of a property and the property itself, rather than writing $\|P\|$ for the property.

[6] Thus, for example, in a domain consisting only of triangles, the algorithm will produce the description 'the blue triangle' to identify a blue triangle, even though 'triangle' is strictly speaking not required to identify the target.

Note that at this point we remain agnostic about how individual salience values are computed, but we will return to this issue later on.

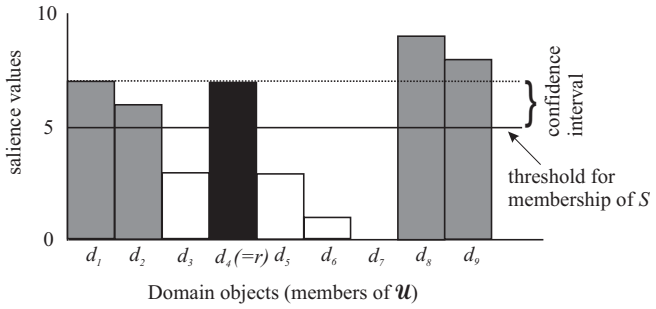

Figure 1: A bar chart depicting for each object in some domain $\mathcal{U}$ the corresponding salience value. The target is represented by a black bar and the other members of the salience property S_r are distinguished by their grey colour.

In *salience-first* accounts, IA is started by initializing C with $S_r(\subseteq \mathcal{U})$ instead of $\mathcal{U}$: the idea is to find a description which distinguishes r from the objects in $\mathcal{U}$ that, given a confidence interval, are at least as salient as r itself. Alternatively, *salience-last* accounts modify iteration step 3: the condition $C = \{r\}$ is replaced by $C \cap S_r = \{r\}$. Thus, at the end of each iteration it is checked whether r is the most salient object which fits the description D. Whereas, for example, Theune (2000) and Deemter and Krahmer (2006) propose salience-first accounts, Kelleher et al. (2005) and Krahmer and Theune (2002) describe salience-last algorithms. The former point out that their approaches are to be preferred on computational grounds; by removing from $\mathcal{U}$ all objects that are not a member of S_r, the algorithm, at each step, has to inspect a smaller C than in any salience-last approach. A further possible reason for preferring salience-first is its cognitive plausibility (Van Deemter and Krahmer mention its 'naturalness', though they do not expand on this). Here we want to draw attention to a novel observation: salience-first and salience-last accounts lead to different empirical predictions.

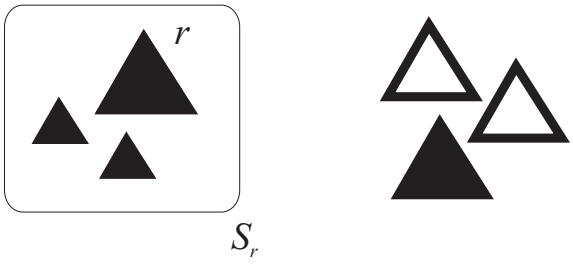

Figure 2: A domain with several triangles. The set of triangles enclosed by the box is the salience property S_r for target r

Consider Figure 2 and let us assume that the attributes are ordered as follows: *shape*, *colour*, *size*.[7] The salience-first approach results in $D = \{big, triangle\}$: C is restricted to the set of salient objects (the ones within the enclosed area). Since all objects are triangles, on the first iteration no property is added to D. On the second iteration, no property is added either (since all salient objects have the same colour). On the third and final iteration, the property *big* is added which distinguishes r from the other objects in C. Finally, $D \cup \{triangle\}$ is returned (since iteration step 3 requires us to add a property from the top-ranked attribute, i.e., from the attribute shape), which can be realized as, for example, 'the big triangle'. Salience-last, in contrast, results in $D = \{black, big, triangle\}$. This is a consequence of the fact that in the second iteration, the test on whether to include *black* is: **a**) Does it include r? *Yes.* **b**) Does it rule out any objects from $\mathcal{U}$ (rather than $S_r(\subseteq \mathcal{U})$)? *Yes, the two white triangles.*

When to point?

In contrast with the accounts of pointing discussed in the introduction of this paper, here we put forward a model for multimodal reference which establishes a direct link between pointing and salience, and more specifically salience-first accounts. The basic ingredients of this approach are:

1. Pointing is a way of making the set of objects that have been pointed at maximally salient.

2. Assuming that the target r is a member of the set of objects that the speaker pointed at, the pointing act causes S_r to be identical with the set of objects that the speaker pointed at.

3. In accordance with the salience-first version of the Incremental Algorithm, S_r (the salience property for r) is used to initialise the context set C, and a description is generated relative to this set. Empirical evidence for the assumption that speakers decide on properties relative to S_r is given Beun and Cremers (1998); they found that if a speaker refers to an object that is part of the focus of attention, s/he usually produces a description that only distinguishes the target from other objects that are part of the focus of attention.

This tells us what the effect of pointing is. We propose that the decision *when* to point is captured by the following rule:

> SALIENCE-BASED POINTING HEURISTIC: If, as a result of pointing, the size of the context set $C = S_r$ for target r can be reduced, then point.

This heuristic may need to be refined for situations where the size of S_r is very small to start with: we may need to add a condition to the rule requiring that $S_r > c$, where c is a constant that has to be determined empirically. Also, the degree

[7] For this particular example, we need the ordering that we provided, but it is straightforward to create examples of the same type based on different orderings.

to which S_r is reduced may play a role. In other words, for both the size of S_r and the degree of its reduction, we may require thresholds.[8]

This account is grounded in the following empirical findings:

1. The decision whether to point is correlated with the salience of the target: pointing is preferred when the target is *not* salient, i.e., when S_r is big relative to the domain $\mathcal{U}$ (Piwek, 2007).

2. When the target is pointed at, on average the number of properties used in the description is smaller (Piwek, 2007).

3. Levelt et al. (1995) and Ruiter (1998) found that the onset of pointing gestures *precedes* their spoken affiliates. This is compatible with the model proposed here, where a speaker *first* decides whether to point and then constructs a verbal description.

Let us compare this approach with the one based on costs advocated by Krahmer and Sluis (2003) (as discussed in the introductory section of this paper). Consider Figure 3. Using the cost assignments provided in Krahmer and Sluis (2003), we can calculate that the optimal description of the target r is 'the small black triangle' (cost 2.25). This description is cheaper than 'this triangle' + pointing (cost 3). Of course, with a different cost assignment (e.g., making verbal properties more expensive and pointing cheaper) the solution changes. More importantly, however, what the cost model does not capture is that pointing is a way to reduce S_r. Compare this with a reference to the target r'. Here we have a small $S_{r'}$ to start with, and pointing may not help from where the speaker is standing: assuming the speaker remains stationary, s/he may only be able to point at a set of objects that is equal to or bigger than $S_{r'}$. The cost-based model does not take these considerations into account.

In the model of Krahmer and Sluis (2003), the decision to point rests on a comparison between the cost of pointing and speaking for the speaker. The cost of pointing is related to the effort involved in making a pointing gesture. In constrast, the current model introduces a salience-based heuristic; speakers point when this helps the speaker quickly construct a referring expression and leads to an expression that can be easily interpreted by the addressee. By choosing to point when this reduces S_r, the speaker makes sure that they only have to identify the target with respect to the smallest possible S_r. An interpreter who knows that speaker acts in this way, can search for the target among the most salient objects in the domain (the ones which his or her attention is focussed on anyway).

[8]One issue that we have factored out of this account concerns the observation reported in Piwek (2007) that some speakers appear to completely refrain from pointing. This suggests that there may be an overriding preference for some speakers not to point.

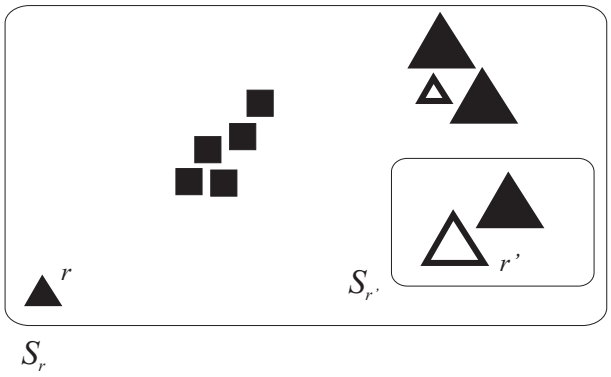

Figure 3: Example of a domain; two targets, r and r', are marked together with their respective salience properties, S_r and $S_{r'}$

Dimensions of salience

So far, we have not dealt with the details of how to compute the salience values that determine S_r. We have suggested that pointing can change salience values. Also, there is ample literature on how verbal reference affects salience. Usually the idea is that the more recent an object was referred to, the more salient it is.[9] In a visually shared domain, spatial relations between objects can also influence salience. In particular, an object that is salient directs attention to itself and the spatial region around it. Consequently, the salience of the objects in its vicinity get a boost - here we will call this implied spatial salience. Beun and Cremers (1998) have found that speakers exploit spatially implied salience in that they usually produce (first-mention) descriptions that only distinguish the target from the most salient object and objects that are spatially implied by (i.e., close to) it. Finally, at the start of a conversation, objects that are central in the scene will be more salient than the objects in the periphery. We will subsume this phenomenon under implied spatial salience: at the beginning of a discourse, the centre of the scene boosts the salience of the objects in its vicinity.[10]

For each of the aforementioned types of salience, we propose to introduce a separate dimension modelled as a function:

- p (pointing dimension),
- v (verbal reference dimension) and
- i (implied spatial dimension).

Each function, when applied to a specific object x returns an integer from $[0-10]$. We also define the aggregate salience value of an object as: $s(x) = max(p(x), v(x), i(x))$. In other words, the overall salience value for an object x is computed

[9]Though the syntactic position of the referring expression also plays a role, e.g., with entities introduced in subject position being more prominent than those introduced in direct object position.

[10]*Cf.* Kelleher et al. (2005).

by taking the maximal value that the salience value has in any of the dimensions.

The dynamics of p, i and v are given by the following equations which relate the dimensions to states (indicated by subscripts):[11]

1. $p_0(x) = v_0(x) = i_0(x) = 0$

2. $p_S(x) = \begin{cases} 10 \text{ if } x \text{ is pointed at between } S-1 \text{ and } S \\ \text{else } 0 \end{cases}$

3. $v_S(x) = \begin{cases} 10 \text{ if condition } \dagger(x) \text{ holds.} \\ v_{S-1}(x) - 1 \text{ if not } \dagger(x) \text{ and} \\ \quad v_{S-1} > 0 \;\&\; \neg\exists y : p_{S-1}(y) = 10 \\ v_{S-1}(x) \text{ if not } \dagger(x) \;\&\; \exists y : p_{S-1}(y) = 10 \\ \text{else } 0 \end{cases}$

4. $i_S(x) = \begin{cases} 8 \text{ if } (\exists y : v_S(y) = 10 \text{ and} \\ \quad x \text{ spatially implies } y) \\ \quad \text{or } (s = 0 \text{ and } sc \text{ spatially implies } x) \\ \text{else } 0 \end{cases}$

Here, *sc* stands for scene centre, and $\dagger(x)$ is an abbreviation of x is referred to between $S-1$ and S.
The equations can be seen at work in Figure 4. This figure depicts a sequence of states for a universe of two objects, d_1 and d_2. Note that in this model states are temporally ordered. Transitions between states can, however, take place in parallel, as long as a transition to a later state is never completed before the transitions to the states preceeding it have been completed.

Equation 1 tells us that in the initial state the salience value for each object in each of the dimensions is 0. Next, equation 2 says that if an object is pointed at between two states ($S-1$ and S), then in the resulting state (S) the salience value for the pointing dimension is set to 10, the highest possible salience value.[12] Equation 3 has four parts which regulate the verbal dimension of the salience value:

- It is set to 10 for an object if the speaker just referred to that object.
- If the speaker did not refer to the object x, the salience value of x is not equal to 0 and no other object was pointed at, then the salience value of x is decreased by 1.
- If the speaker did not refer to the object x, but pointed to some other object y, then the salience value of x does not change. This means that if a speaker refers to an object by means of a multimodal referring act (pointing + a verbal reference), then the decay of the salience of all other objects is caused by the pointing act (and not the subsequent reference). Without this clause, all other objects would be decreased twice by 1 in the course of a multimodal referring act (as a result of the pointing act *and* then again the verbal reference). This would go against the idea that a multimodal referring act is no different from a unimodal act in terms of the update on salience values of objects that were not referred to.
- Finally, if none of the aforementioned conditions holds, the salience value in the verbal dimension is set to 0.

Equation 4 spells out how spatially implied objects are assigned a salience value of 8: an object that is next to a maximally salient object receives the salience value 8, and also at the beginning of a discourse, any objects that are close to the centre of the scene are reset to salience value 8.[13]

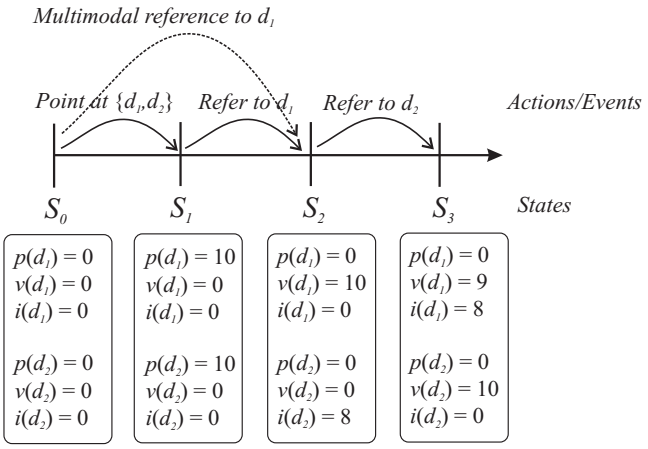

Figure 4: Example of how salience values change as a result of pointing and reference. p, v and i stand for the three dimensions of salience: the pointing, verbal reference, and implied spatial dimension.

Let us briefly discuss our assumptions about the pointing act itself. A pointing act is viewed as raising the salience of a set of objects (though in the limiting case this set can be a singleton set). It is a set because even though a speaker may intend to single out a specific object, usually this is not possible. The speaker aims for the location of the object. As this object is further away, the location which the speaker may be pointing at becomes less and less definite because both speaker and addressee will be increasingly uncertain about which points

[11] Our account is restricted to modelling the trajectories of the salience values of objects in a shared domain of conversation. We have not attempted to integrate it with an account of the informational content that is exchanged during the conversation. We view it as a future project to integrate the current model with, for example, Discourse Representation Theory (Kamp & Reyle, 1993) or Situation Semantics (Barwise & Perry, 1983). Some results on integrating Situation Semantics with attentional state have already been obtained by Poesio (1993).

[12] Often a pointing act will not unambiguously be directed at a single object. In that case, all the objects that the speaker is pointing at are affected by this equation (i.e., their salience value is set to 10).

[13] The choice for the value 8 needs to be empirically validated. The idea behind this is that spatially implied objects are less salient than the most recently referred to object, but more salient than objects that were referred to about two references ago.

in space the line extended from the speaker's index finger intersects with. If there are many objects in the vicinity of this line, this will lead to uncertainty about which object the speaker pointed at. Consequently, a pointing act will typically identify a set of objects that are potentially the target of the pointing act.[14] A second important assumption we make is that the speaker is stationary. Of course, if a speaker were to move sufficiently close to the target, s/he could make sure that the pointing act only identifies the target. In some situations, this may be the appropriate thing to do. For now, we simply assume that the speaker is not allowed to move. If s/he were allowed to move, it might be necessarry to factor in the cost of moving against that of pointing less precisely, thus possibly introducing some sort of cost-based calculation along the lines of Krahmer and Sluis (2003) and Sluis and Krahmer (2007).

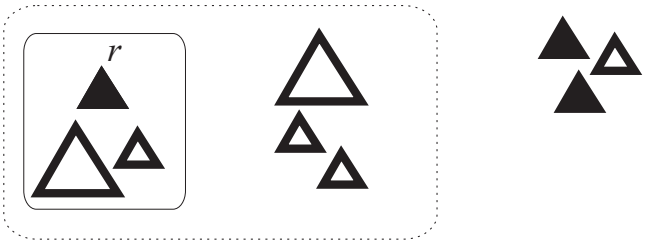

Figure 5: The dotted line indicates the set of objects that the speaker pointed at. The solid line includes the target *r* and the objects that are spatially implied by the target.

We have proposed a model that keeps track of the salience values in the three dimensions separately. We conclude this section by showing that, in particular, spatially implied salience and pointing salience need to be kept apart. Let us first explain the intuition behind this. The idea is that pointing identifies a set of potential targets. Subsequently, the verbal reference identifies the actual target. At that point the candidates in the pointing set are no longer relevant; they only needed to be taken into account as long as there existed uncertainty about the interpretation of the pointing act. Thus, intuitively, there is a difference between the set of objects that the speaker pointed at and the set of objects that are spatially implied by the target. The best way to illustrate the difference is to examine an example where the two diverge. Take Figure 5. Suppose the speaker points at the objects that are enclosed by the dotted line and says 'the black triangle' thereby identifying the target *r*. Now, assume that the next thing the speaker says is 'the big white triangle'. In this case, our model predicts that the speaker is talking about the triangle that is located immediately below *r*. However, if we had not distinguished between the *p* and *i* dimensions, and for example assumed that *i* was identical to *p*, then the utterance of 'the big white triangle' would have been ambigous between the two big white triangles enclosed by the dotted line. We conjecture that the latter prediction is incorrect and intend to verify this empirically.

[14] See Kranstedt, Lücking, Pfeiffer, Rieser, and Wachsmuth (2006) for an empirical study into how to assign extensions to pointing acts.

Conclusions

This paper started by distinguishing between salience-first and salience-last approaches to integrating salience with the generation of referring expressions. We demonstrated that the approaches differ not only in computational efficiency, but also in empirical predictions. We then proceeded to describe a model of multimodal reference. The proposal follows the insight from the study of human pointing gestures that pointing is primarily a means for changing the salience of objects. Our account is framed in terms of a salience-first algorithm. We proposed a salience-based pointing heuristic which suggest that speakers point when they can thereby reduce the number of other objects in the domain from which the target needs to be distinguished. The proposal is grounded in a number of empirical findings about human multimodal referring acts and will hopefully provide a fruitful starting point for further experimental studies into production of multimodal referring acts.

Acknowledgments

I would like to thank the three ananonymous reviewers for PRE-CogSci 2009 and my colleague Richard Power for helpful feedback on a draft of this paper.

References

Ariel, M. (1990). *Accessing noun-phrase antecedents*. London: Routledge.

Barwise, J. (1989). On the Model Theory of Common Knowledge. In *The Situation in Logic* (p. 201-220). Stanford, CA.: CSLI.

Barwise, J., & Perry, J. (1983). *Situations and Attitudes*. Cambridge, MA: MIT Press.

Beun, R., & Cremers, A. (1998). Object reference in a shared domain of conversation. *Pragmatics & Cognition*, *6*(1/2), 121-152.

Butterworth, B., & Hadar, U. (1989). Gesture, speech and computational stages: A reply to McNeill. *Psychological Review*, *96*, 1–47.

Butterworth, G. (2003). Pointing is the royal road to language for babies. In S. Kita (Ed.), *Pointing: Where Language, Culture and Cognition Meet* (p. 9-34). Mahwah, NJ: Lawrence Erlbaum Associates.

Choumane, A., & Siroux, J. (2008). Knowledge and Data Flow Architecture for Reference Processing in Multimodal Dialogue Systems. In *2008 Conference on Multimodal Interfaces (ICMI'08)*. Crete, Greece.

Claassen, W. (1992). Generating referring expressions in a multimodal environment. In R. D. et al. (Ed.), *Aspects of Automated Natural Language Generation*. Berlin: Springer Verlag.

Clark, H. (1996). *Using language*. Cambridge: Cambridge University Press.

Clark, H., Schreuder, R., & Buttrick, S. (1983). Common ground and the understanding of demonstrative reference. *Journal of Verbal Learning and Verbal Behavior*, *22*, 1-39.

Dale, R., & Reiter, E. (1995). Computational interpretation of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, *19*(8), 233–263.

Deemter, K. van, & Krahmer, E. (2006). Graphs and Booleans: on the generation of referring expressions. In H. Bunt & R. Muskens (Eds.), *Computing meaning* (Vol. 3). Dordrecht: Kluwer.

Feyereisen, P. (2007). How do gesture and speech production synchronise? *Current psychology letters*, *2*(22), 2-12.

Goodwin, C. (2003). Pointing as situated practice. In S. Kita (Ed.), *Pointing: Where Language, Culture and Cognition Meet* (p. 217-241). Mahwah, NJ: Lawrence Erlbaum Associates.

Grosz, B., & Sidner, C. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, *12*(3).

Kahneman, D. (2003). A perspective on judgement and choice: Mapping bounded rationality. *American Psychologist*, *58*(9), 697–720.

Kamp, H., & Reyle, U. (1993). *From Discourse to Logic: Introduction to Modeltheoretic Semantics for Natural Language, Formal Logic and Discourse Representation Theory*. Dordrecht: Kluwer Academic Publishers.

Kelleher, J., Costello, F., & Genabith, J. van. (2005). Dynamically structuring, updating and interrelating representations of visual and linguistic discourse context. *Artificial Intelligence*, *167*, 62-102.

Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, *48*, 16-32.

Krahmer, E., & Sluis, I. van der. (2003). A new model for the generation of multimodal referring expressions. In *Proceedings European Workshop on Natural Language Generation (ENLG2003).* Budapest, Hungary.

Krahmer, E., & Theune, M. (2002). Efficient context-sensitive generation of referring expressions. In K. van Deemter & R. Kibble (Eds.), *Information Sharing* (p. 223-264). Stanford University: CSLI.

Kranstedt, A., Lücking, A., Pfeiffer, T., Rieser, H., & Wachsmuth, I. (2006). Deictic object reference in task-oriented dialogue. In G. Rickheit & I. Wachsmuth (Eds.), *Situated communication* (p. 155-208). Mouton de Gruiter.

Krauss, R., Chen, Y., & Chawla, P. (1996). Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 28, p. 389-450). Academic Press.

Lester, J., Voerman, J., Towns, S., & Callaway, C. (1999). Deictic Believability: Coordinating gesture, locomotion, and speech in lifelike pedagogical agents. *Applied Artificial Intelligence*, *13*(4-5), 383414.

Levelt, W. (1989). *Speaking: From Intention to Articulation*. Cambridge, Massachusetts: The MIT Press.

Levelt, W., Richardson, G., & Heij, L. (1995). Pointing and voicing in deictic expressions. *Journal of Memory and Language*, *24*, 133-164.

McNeill, D. (2005). *Gesture and Thought*. Chicago: University of Chicago Press.

Piwek, P. (2007, January). Modality choice for generation of referring acts: Pointing versus describing. In *Proceedings of workshop on multimodal output generation (mog 2007)* (pp. 129–139). Aberdeen, Scotland.

Piwek, P., Beun, R., & Cremers, A. (2008). 'Proximal' and 'Distal' in language and cognition: evidence from deictic demonstratives in Dutch. *Journal of Pragmatics*, *40*(4), 694-718.

Poesio, M. (1993). A Situation-Theoretic Formalization of Definite Description Interpretation in Plan Elaboration Dialogues. In P. Aczel, D. Israel, Y. Katagiri, & S. Peters (Eds.), *Situation Theory and its Applications* (Vol. 3, p. 339-374). CSLI.

Ruiter, J. de. (1998). *Gesture and speech production*. Unpublished doctoral dissertation, Max Planck Institute, Nijmegen.

Ruiter, J. de. (2000). The production of gesture and speech. In D. McNeill (Ed.), *Language and Gesture* (p. 284-311). Cambridge: Cambridge University Press.

Schelling, T. (1960). *The strategy of conflict*. Cambridge, Mass.: Harvard University Press.

Sluis, I. van der, & Krahmer, E. (2001). Generating Referring Expressions in a Multimodal Context: An empirically motivated approach. In *Selected Papers from the 11th CLIN Meeting.* Amsterdam: Rodopi.

Sluis, I. van der, & Krahmer, E. (2007). Generating multimodal referring expressions. *Discourse Processes*, *44*(3), 145–174.

Theune, M. (2000). *From Data to Speech: Language Generation in Context*. Unpublished doctoral dissertation, Eindhoven University of Technology.